PB86-208659

Clustering Criteria and Multivariate Normal Mixtures

North Carolina Univ. at Chapel Hill

Prepared for

Health Effects Research Lab.
Research Triangle Park, NC

1981

U.S. Department of Commerce
National Technical Information Service
NTIS

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO.	2.	PB86-208659
EPA/600/J-81/698		

4. TITLE AND SUBTITLE
CLUSTERING CRITERIA AND MULTIVARIATE NORMAL MIXTURES
(Journal Version)

5. REPORT DATE
1981

6. PERFORMING ORGANIZATION CODE

7. AUTHOR(S)
M. J. Symons

8. PERFORMING ORGANIZATION REPORT NO.

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Department of Biostatistics
UNC, Chapel Hill, N.C.

10. PROGRAM ELEMENT NO.

11. CONTRACT/GRANT NO.

12. SPONSORING AGENCY NAME AND ADDRESS
Health Effects Research Laboratory-RTP, NC
Office of Research and Development
U.S. Environmental Protection Agency
Research Triangle Park, NC 27711

13. TYPE OF REPORT AND PERIOD COVERED

14. SPONSORING AGENCY CODE
EPA/600/11

15. SUPPLEMENTARY NOTES
John P. Creason, BD/HERL REF: Biometrics, 37:35-43, March 1981

16. ABSTRACT

New clustering criteria for use when a mixture of multivariate normal distributions is an appropriate model are presented. They are derived from maximum likelihood and Bayesian approaches corresponding to different assumptions about the covariance matrices of the mixture components. Two of the criteria are modifications of the determinant of the within-groups sum-of-squares criterion of Friedman and Rubin (1967, Journal of the American Statistical Association 63, 1159-1178); these criteria appear to be more sensitive to disparate cluster sizes. Two others are appropriate for different-shaped clusters.

The practical aspects of these criteria, and of another one studied by Maronna and Jacovkis (1974, Biometrics 30, 499-505) for heterogeneous covariance matrices, are outlined. An example involving the separation of two types of diabetic patients from normal subjects, each group having a distinct covariance structure, is given. The results with the three criteria appropriate for different-shaped clusters were comparable to one another and preferable to those obtained with the three criteria for similar-shaped clusters. Results obtained for the example with two additional clustering procedures are presented.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS (This Report)	21. NO. OF PAGES
RELEASE TO PUBLIC	UNCLASSIFIED	10
	20. SECURITY CLASS (This page) UNCLASSIFIED	22. PRICE

EPA Form 2220-1 (Rev. 4-77)

REPRODUCED BY
NATIONAL TECHNICAL INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

NTIS is authorized to reproduce and sell this report. Permission for further reproduction must be obtained from the copyright owner.

EPA/600/J-81/698
JOURNAL ARTICLE

BIOMETRICS 37, 35–43
March 1981

Clustering Criteria and Multivariate Normal Mixtures

M. J. Symons

Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina 27514, U.S.A.

SUMMARY

New clustering criteria for use when a mixture of multivariate normal distributions is an appropriate model are presented. They are derived from maximum likelihood and Bayesian approaches corresponding to different assumptions about the covariance matrices of the mixture components. Two of the criteria are modifications of the determinant of the within-groups sum-of-squares criterion of Friedman and Rubin (1967, *Journal of the American Statistical Association* **63**, 1159–1178); these criteria appear to be more sensitive to disparate cluster sizes. Two others are appropriate for different-shaped clusters.

The practical aspects of these criteria, and of another one studied by Maronna and Jacovkis (1974, *Biometrics* **30**, 499–505) for heterogeneous covariance matrices, are outlined. An example involving the separation of two types of diabetic patients from normal subjects, each group having a distinct covariance structure, is given. The results with the three criteria appropriate for different-shaped clusters were comparable to one another and preferable to those obtained with the three criteria for similar-shaped clusters. Results obtained for the example with two additional clustering procedures are presented.

1. Introduction and Model

Several authors, including Wolfe (in Research Memoranda SRM 68-2 and SRM 69-17 of the U.S. Naval Personnel Research Activity, San Diego, California, 1967 and 1969), Day (1969) and Binder (1978), have considered clustering problems in which a mixture of multivariate normals is used as a statistical model. With the number of components in the mixture specified as G, the clustering problem may be formulated as the estimation of the mixture component origin of each of the n p-variate observations, $\mathbf{y}_i$, where i indexes the sample.

More specifically, the density of $\mathbf{y}_i$ is

$$f(\mathbf{y}_i \mid \pi_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) = \sum_{g=1}^{G} \pi_g N_p(\mathbf{y}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \tag{1}$$

where the terms π_g are the mixing parameters, each required to be positive, their sum being unity, while the notation $N_p(\mathbf{y}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ denotes that $\mathbf{y}_i$ is distributed as a p-variate normal with mean vector $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$.

The unknown mixture component origin for $\mathbf{y}_i$ is denoted by z_i. From the mixture model, we have z_i equals g, or equivalently $\mathbf{y}_i$ comes from the gth component, with probability π_g. The clustering problem then can be viewed as the problem of estimating the n values of z_i.

Key words: Homogeneous and heterogeneous groups; Variable metric.

2. Clustering Criteria

For both the maximum likelihood and Bayesian approaches, the likelihood of the data is required in order to estimate the n components in the vector $\mathbf{z} = (z_1, z_2, \ldots, z_n)$. If the matrix $\mathbf{Y}$ denotes the n observations $\mathbf{y}_i$, and $\boldsymbol{\theta}$ the vector of parameters $(\pi_1, \ldots, \pi_G, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_G)$, then the likelihood of the data is given by

$$L(\mathbf{Y} \mid \boldsymbol{\theta}, \mathbf{z}) = \prod_{g=1}^{G} (\pi_g^{n_g} |\boldsymbol{\Sigma}_g|^{-\frac{1}{2}n_g}) \exp\left\{-\tfrac{1}{2} \sum_{g=1}^{G} \sum_{C_g} (\mathbf{y}_i - \boldsymbol{\mu}_g)' \boldsymbol{\Sigma}_g^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_g)\right\}, \tag{2}$$

where C_g is the collection of $\mathbf{y}_i$ observations with z_i equal to g, and n_g is the number of observations in C_g.

Several points about this likelihood are noteworthy. (i) It is conditional on the parameters $\boldsymbol{\theta}$ and a specified vector $\mathbf{z}$, which allocate each $\mathbf{y}_i$ to one of the G components in the mixture (1). (ii) As there are G^n possible allocations, approximate search routines are required to provide the optimal assignment of the observations to the groups and thereby to determine the clusters $C_1, C_2, \ldots, C_G$. (iii) The product of the $\pi_g^{n_g}$ terms over the G groups is the likelihood of *number* of observations assigned to each group. The rest of $L(\mathbf{Y} \mid \boldsymbol{\theta}, \mathbf{z})$ is the likelihood of *which* n_g observations, for $g = 1, \ldots, G$, are allocated to the respective groups. (iv) This latter portion of likelihood (2) is equivalent to the likelihood given by Scott and Symons (1971), which does not presume a probabilistic structure for the event of $\mathbf{y}_i$ coming from the gth component.

The maximum likelihood (ML) approach determines the ML estimate, $\hat{\mathbf{z}}$, of $\mathbf{z}$ as the allocation that maximizes (2). The parameters are replaced by their ML estimates, given an allocation of the n observations $\mathbf{y}_i$ to the G components. These ML estimates are the standard ones, namely:

$$\hat{\pi}_g = n_g/n, \tag{3}$$

$$\hat{\boldsymbol{\mu}}_g = \bar{\mathbf{y}}_g = \frac{1}{n_g} \sum_{C_g} \mathbf{y}_i \tag{4}$$

and

$$\hat{\boldsymbol{\Sigma}}_g = \frac{1}{n_g} \mathbf{W}_g = \frac{1}{n_g} \sum_{C_g} (\mathbf{y}_i - \bar{\mathbf{y}}_g)(\mathbf{y}_i - \bar{\mathbf{y}}_g)', \tag{5}$$

for $g = 1, \ldots, G$.

The Bayesian approach, discussed generally by Lindley (1966), requires specification of a prior distribution $p(\boldsymbol{\theta})$, for $\boldsymbol{\theta}$. The priors of Jeffreys (1961) were utilized in order to delineate simply the bounds of the parameter space. As the parameters $\boldsymbol{\theta}$ are not of central interest, the likelihood (2) is averaged over the prior $p(\boldsymbol{\theta})$. Specifically,

$$L(\mathbf{Y} \mid \mathbf{z}) = \int L(\mathbf{Y} \mid \boldsymbol{\theta}, \mathbf{z}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \tag{6}$$

where the integration is over the parameter space of $\boldsymbol{\theta}$. Geisser (1966) referred to the normalization of (6) as the predictive distribution of $\mathbf{Y}$. The mode of (6) is taken as the Bayes estimate, $\tilde{\mathbf{z}}$, of the optimal allocation.

Clustering criteria for these two approaches are presented for the cases (i) when the mixture components in (1) have unknown covariance matrices that are assumed to be homogeneous, and (ii) when these matrices are not necessarily equal. The component origin of each observation is presumed unknown, and none of the parameters in (1) is presumed known. For other cases, and situations with some parameters specified and/or

with the compone
Symons (1971), S
Biostatistics. Univ
references also co

2.1 *Covariance M*

When $\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$ for
likelihood over G
and

replace the param
the n observation

where $|\mathbf{W}|$ is the c
The Bayesian a

$p(\boldsymbol{\theta}) =$

to define the para
Cornfield (1963).
in (6). The Bayesi
groups that minim

The ML criteri
the within-groups

2.2 *Covariance M*

When the covaria
$\hat{\mathbf{z}}$, is that partitio

Criterion (11) diff
(5), are required
mixture. In orde
least $p - 1$ obser
The Bayesian
independently fo
(6) and is equiva

$$\sum_{g=1}^{G} ((n_g - 1) \ln |\mathbf{W}
$$

with the component origin of some observations known, see Geisser (1966), Scott and Symons (1971), Symons (Institute of Statistics Mimeo Series No. 880, Department of Biostatistics. University of North Carolina at Chapel Hill, 1973) and Binder (1978). These references also contain more details for the Bayesian approach to this problem.

2.1 *Covariance Matrices Homogeneous*

When $\Sigma_g = \Sigma$ for $g = 1, \ldots, G$, and Σ is unknown, the ML approach is to maximize the likelihood over G^n possible allocations. For each allocation $\mathbf{z}$, the ML estimators (3), (4) and

$$\hat{\Sigma} = \frac{1}{n}\mathbf{W} = \frac{1}{n}\sum_{g=1}^{G}\mathbf{W}_g \tag{7}$$

replace the parameters $\boldsymbol{\theta}$. The ML optimal allocation, $\hat{\mathbf{z}}$, is equivalent to the partition of the n observations into G groups which minimizes the criterion

$$n \ln |\mathbf{W}| - 2\sum_{g=1}^{G} n_g \ln n_g, \tag{8}$$

where $|\mathbf{W}|$ is the determinant of the within-groups sum of squares.

The Bayesian approach utilizes a vague prior, namely

$$p(\boldsymbol{\theta}) = p(\pi_1, \ldots, \pi_G)p(\mu_1, \ldots, \mu_G \mid \Sigma)p(\Sigma) \propto \left(\prod_{g=1}^{G}\pi_g\right)^{-1} |\Sigma|^{-\frac{1}{2}(p+1)}, \tag{9}$$

to define the parameter space. The prior on Σ is of the general form used by Geisser and Cornfield (1963). The product of (2) and (9) is averaged over the parameter space of $\boldsymbol{\theta}$, as in (6). The Bayesian optimal allocation, $\tilde{\mathbf{z}}$, is equivalent to the partition of the data into G groups that minimizes the criterion

$$(n-G)\ln|\mathbf{W}| + \sum_{g=1}^{G}\{p \ln n_g - 2\ln\Gamma(n_g)\}. \tag{10}$$

The ML criterion (8) and Bayes criterion (10) are modifications of the determinant of the within-groups sum of squares, a criterion proposed by Friedman and Rubin (1967).

2.2 *Covariance Matrices Unequal and Unknown*

When the covariance matrices may differ from group to group, the ML optimal allocation, $\hat{\mathbf{z}}$, is that partition of the n observations into G groups, which minimizes the criterion

$$\sum_{g=1}^{G} n_g \ln|\mathbf{W}_g| - 2\sum_{g=1}^{G} n_g \ln n_g. \tag{11}$$

Criterion (11) differs from (8) in that the within-groups sum-of-squares matrices, $\mathbf{W}_g$ from (5), are required to estimate an ellipsoidal shape for each of the G components in the mixture. In order that none of the $\mathbf{W}_g$ should be singular, each cluster must contain at least $p+1$ observations.

The Bayesian approach for this case utilizes vague prior information as in (9), but independently for each of the G matrices Σ_g. The optimal allocation, $\tilde{\mathbf{z}}$, is determined by (6) and is equivalent to the partition that minimizes the criterion

$$\sum_{g=1}^{G}\left((n_g-1)\ln|\mathbf{W}_g| + p\ln n_g - p(n_g+p)\ln 2 - 2\left[\ln\Gamma(n_g) + \sum_{i=1}^{p}\ln\Gamma\left\{\tfrac{1}{2}(n_g+p+1-i)\right\}\right]\right). \tag{12}$$

This Bayes criterion is considerably more complicated than the ML criterion (11). The additional complexity stems from the normalization by (6) of a Wishart density for each of the G components in (1).

One additional criterion for different-shaped clusters is presented and compared with (11) and (12) in the next section. Maronna and Jacovkis (1974) compared several multivariate clustering procedures with variable metrics. Their study included numerical experiments using real data and Monte Carlo data simulating mixtures of G normal p-variate populations. The variable metric, based upon the suggestions of Chernoff (1970) and especially those of Rohlf (1970), was the only metric 'for which the clustering method had reasonable properties'. It was shown to be equivalent to the clustering criterion

$$\sum_{g=1}^{G} |\mathbf{W}_g|^{1/p} \tag{13}$$

when translated from the notation of Maronna and Jacovkis, and is based upon the within-cluster covariance matrix normalized to have unit determinant. By showing this equivalence between a metric clustering procedure and the optimization of a clustering criterion, or 'uncertainty functional', Maronna and Jacovkis noted an important connection between quite different approaches to cluster analysis.

3. Sketch of Empirical Performances

A thorough comparison of the performances of various approaches to clustering should cover several factors, including the number of groups, number of variables, shape of the clusters, similarity of shape from cluster to cluster, relative sizes of the clusters, separation of the clusters, and total sample size available.

The criteria presented in the previous section are determined on the assumption that the number of groups is specified. Inference as to the number of groups is usually required in a clustering problem; Lee (1979) provided an approach, and Everitt (1979) gave a general review of efforts on this basic question. The following subsections provide a sketch of the performance of the criteria when the number of groups is known. Only selected combinations of the remaining factors that bear on the sensitivity and specificity of the criteria are considered below. In addition, results with NORMIX (cf. the 1967 report by Wolfe referred to in §1) and with the SAS (1979) CLUSTER procedure are given for comparison with the criteria of §2 in an example containing three heteroscedastic clusters. A summary of these practical considerations is given in §3.3.

3.1 *Performance with Homogeneous Groups*

When the cluster sizes are nearly equal, the method employing the determinant of the within-groups sum-of-squares matrix performs very well. When the group sizes are disparate, it has been observed by Scott and Symons (1971) and by Binder (1978) that the $|\mathbf{W}|$ criterion tends to favor partitions of equal size. In this case, the ML criterion (8) and the Bayes criterion (10) perform better, but may tend to overemphasize slightly the size of the larger groups, as was shown in the author's 1973 report.

Some preliminary empirical experience of the author with criteria (11), (12) and (13), which are appropriate for different-shaped clusters, suggests that these criteria may produce clusters of different shapes and sizes when presented with homogeneous-shaped clusters that are close together.

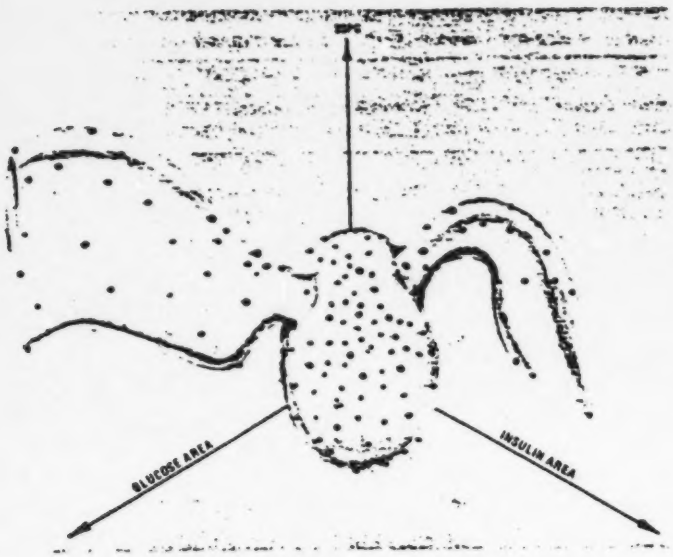

Figure 1. Artist's representation of the data of Reaven and Miller (1979, p. 19), as seen in three dimensions.

3.2 *Example and Performance with Heterogeneous Groups*

The relationship between chemical diabetes and overt diabetes in 145 nonobese adult subjects was examined by Reaven and Miller (1979). The degree of glucose intolerance, insulin response to oral glucose, and insulin resistance was determined in normal subjects and in patients with nonketotic diabetes. The three-dimensional shape of the data set was that of a 'boomerang with two wings and a fat middle', and is reproduced in Fig. 1. The two wings were interpreted as representing patients with chemical diabetes and with overt diabetes, respectively, while the spherical middle corresponds to normal subjects.

The heterogeneity of the parts of this three-dimensional shape and their nonellipsoidal nature provide a practical test for the criteria presented in §2. Reaven and Miller (1979) utilized the determinant of the within-groups sum of squares as their clustering criterion, supplemented by typical central values for groups from an earlier set of 125 patients, that were similar to the groups presented in Fig. 1. This *a priori* information was not used in the analyses reported here, as the capability of the clustering criteria alone was of primary interest.

These criteria were all options in an approximate routine constructed by McRae (1971) and slightly modified. After determining, from among 32 randomly-generated partitions, the partition that corresponds to a minimum value of the selected criterion, the routine produces a relative minimum for that criterion in the sense that any reassignment of one observation results in a larger value. This algorithm does not guarantee an absolute minimum over the G^n possible allocations; in fact, numerous local minima were often found. Because of this feature, several runs were made with variations in the algorithm. Also, the partitions in the neighbourhood of the clinical classification, based upon the results of the glucose tolerance test and reported by Reaven and Miller (1979), were examined.

Table 1

Clustering of Reaven's and Miller's trivariate diabetes data into three groups

Classification scheme	Group* Normals (76, 0, 0)	Chemical diabetes (0, 36, 0)	Overt diabetes (0, 0, 33)	Criterion value
Variant of $\|W\|$ with *a priori* cluster locations specified; Reaven and Miller (1979)	(73, 10, 1)	(3, 26, 6)	(0, 0, 26)	Not given
Determinant of within-groups sum of squares, $\|W\|$; see (7), and Friedman and Rubin (1967)	1†: (76, 0, 0)	(0, 36, 0)	(0, 0, 33)	0.2217×10^{19}
	2: (73, 17, 3)	(3, 19, 4)	(0, 0, 26)	0.7416×10^{18}
	3: (73, 17, 3)	(3, 19, 4)	(0, 0, 26)	0.7416×10^{18}
	4: No other local minima found			
Maximum likelihood modification of $\|W\|$; see (8)	1: (76, 0, 0)	(0, 36, 0)	(0, 0, 33)	0.4978×10^{4}
	2: (75, 30, 6)	(1, 6, 1)	(0, 0, 26)	0.4750×10^{4}
	3: (75, 30, 6)	(1, 6, 1)	(0, 0, 26)	0.4750×10^{4}
	4: (76, 36, 7)	(0, 0, 11)	(0, 0, 15)	0.4754×10^{4}
Bayesian modification of $\|W\|$; see (10)	1: (76, 0, 0)	(0, 36, 0)	(0, 0, 33)	0.5182×10^{4}
	2: (75, 30, 6)	(1, 6, 1)	(0, 0, 26)	0.4951×10^{4}
	3: (75, 30, 6)	(1, 6, 1)	(0, 0, 26)	0.4951×10^{4}
	4: (76, 36, 7)	(0, 0, 11)	(0, 0, 15)	0.4953×10^{4}
Metric criterion with $\|W_e\|$; see (13), and Maronna and Jacovkis (1974)	1: (76, 0, 0)	(0, 36, 0)	(0, 0, 33)	0.8141×10^{6}
	2: (73, 10, 0)	(3, 26, 7)	(0, 0, 26)	0.6123×10^{6}
	3: (73, 10, 0)	(3, 26, 7)	(0, 0, 26)	0.6123×10^{6}
	4: (73, 12, 1)	(3, 24, 6)	(0, 0, 26)	0.6134×10^{6}
Maximum likelihood criterion with $\|W_e\|$; see (11)	1: (76, 0, 0)	(0, 36, 0)	(0, 0, 33)	0.4225×10^{4}
	2: (63, 0, 0)	(13, 30, 2)	(0, 6, 31)	0.4128×10^{4}
	3: (63, 0, 0)	(13, 30, 2)	(0, 6, 31)	0.4128×10^{4}
	4: (68, 5, 0)	(7, 25, 2)	(1, 6, 31)	0.4130×10^{4}
Bayesian criterion with $\|W_e\|$; see (12)	1: (76, 0, 0)	(0, 36, 0)	(0, 0, 33)	0.3095×10^{4}
	2: (73, 9, 0)	(3, 27, 7)	(0, 0, 26)	0.3005×10^{4}
	3: (69, 9, 0)	(7, 23, 4)	(0, 4, 29)	0.300444×10^{4}
	4: (72, 9, 0)	(4, 23, 2)	(0, 4, 31)	0.300457×10^{4}

* This division into three groups of the 145 patients, based upon the glucose tolerance test, is given in Reaven and Miller (1979).

† The four table entries are: 1: Classification and criterion value for the clinical classification. 2: Classification with minimum criterion value found by starting at the clinical classification. 3 and 4: Classification with minimum criterion value for the two smallest criteria values produced by McRae's (1971) program.

Table 1 contains the results obtained using each of the following criteria: the determinant of the within-groups sum of squares; the ML and the Bayesian modifications of the same, (8) and (10) respectively; the ML (11) and the Bayes (12) criteria for heterogeneous clusters; the variable metric of Maronna and Jacovkis, which is equivalent to criterion (13). The notation (73, 10, 1), (3, 26, 6), (0, 0, 26) gives the resulting composition of the three clusters found by Reaven and Miller, relative to the clinical classification given at the head of Table 1. Specifically, the first cluster (73, 10, 1) is composed of 73 normals, 10 chemical diabetics and one overt diabetic. The second cluster is composed of three

normals, 26 chemical d
third cluster.

There are four table e
the clinical classificatio
found by starting at t
partitions correspondin
program.

The best solution is u
Fig. 1, the clinical classi
results from 125 earli
produced reasonable clu

For these heterogene
riate for homogeneous
obtained with the ML (
1, can be very poor. T
different-shaped cluster

In addition these dat
1967 report. An outlin
program, with homoger
(0, 0, 26). These are id
(10) modification of th
allowed, the NORMIX
comparable to those
heterogeneous covarian
the user's choice of cov

For further comparisc
these data. This proce
defines distance betwee
one cluster and an obs
procedure, with or with
is comparable to the se
then, may produce poo

3.3 *Summary of Practi*

The paper by Everitt (
that are of concern to u
comparison of these m

There seems to be n
other approaches to clus
of some empirical resul
depends upon a knowl
analysis, the only reas
several approaches ar
dimensional scatter dia

A note of caution is
many data sets. For e
third-best obtained w
the second-best solutic

normals, 26 chemical diabetics and six overt diabetics, while 26 overt diabetics form the third cluster.

There are four table entries for each of the six criteria. The first is the criterion value for the clinical classification. The second is the clustering with a minimum criterion value found by starting at the clinical classification. The third and fourth entries are the partitions corresponding to the two smallest criterion values located by McRae's (1971) program.

The best solution is unknown for this data set. However, the combined impression from Fig. 1, the clinical classification and the results obtained by Reaven and Miller utilizing the results from 125 earlier patients suggests that the criteria (11), (12) and (13) have produced reasonable clusters. These criteria are appropriate for different-shaped clusters.

For these heterogeneous-shaped groups, the results obtained with the criteria appropriate for homogeneous clusters are not good. The $|\mathbf{W}|$ result is better than the clusters obtained with the ML (8) or the Bayes (10) modification of $|\mathbf{W}|$, which, according to Table 1, can be very poor. These three criteria, then, may produce misleading results when different-shaped clusters are present.

In addition these data were analyzed by the NORMIX program described in Wolfe's 1967 report. An outline of this procedure is provided in §4. The results from Wolfe's program, with homogeneous covariance matrices specified, were: (75, 30, 6), (1, 6, 1) and (0, 0, 26). These are identical to the best results obtained by the ML (8) and the Bayes (10) modification of the $|\mathbf{W}|$ criteria. When heterogeneous covariance matrices were allowed, the NORMIX program produced the solution: (69, 7, 0), (7, 23, 2) and (0, 6, 31), comparable to those from criteria (11), (12) and (13) which are appropriate for heterogeneous covariance structures. Good results with NORMIX then can depend upon the user's choice of covariance structure.

For further comparison, the SAS (1979) CLUSTER procedure was also used to analyze these data. This procedure is hierarchical in nature, employs a Euclidean metric and defines distance between two clusters as the maximum distance between an observation in one cluster and an observation in the other cluster. The solution with this CLUSTER procedure, with or without standardization, was: (76, 36, 7), (0, 0, 9) and (0, 0, 17), which is comparable to the second-best solutions found by criteria (8) and (10). This approach, then, may produce poor results when different-shaped clusters are present.

3.3 *Summary of Practical Aspects*

The paper by Everitt (1979) provides a comprehensive overview of the practical issues that are of concern to users of cluster analysis methods. It stresses the critical review and comparison of these methods, aspects worthy of emphasis.

There seems to be no simple recommendation to guide the users of these criteria or of other approaches to cluster analysis. However, the general impression from the above sketch of some empirical results is that the choice of the most appropriate criterion or approach depends upon a knowledge of the structure of the data set. Since this is the object of the analysis, the only reasonable *a priori* suggestion is to compare results obtained from several approaches and the impressions from visual plots. The plots may be two-dimensional scatter diagrams, or three-dimensional ones such as Fig. 1.

A note of caution is also in order. Multiple local minima are usually encountered with many data sets. For example, the solution: (68, 3, 0), (8, 25, 4) and (0, 8, 29) was the third-best obtained with a criterion (11) value of 0.4131×10^4, nearly identical to the second-best solution with (11). The solution: (73, 9, 0), (3, 27, 7) and (0, 0, 26) was

the third-best obtained with a criterion (12) value of 0.300528×10^4, very close to the second-best solution with (12). Friedman and Rubin (1967) also noted this difficulty. Improved software may alleviate some of these problems, but the existence of multiple local minima with close or overlapping groups should be expected.

The available sample size is also a factor to be considered. How large this sample size is depends upon the number of groups, dimension of the data and the relative sizes of the groups. The effect of sample size on the performance of the criteria of §2 and on the NORMIX program is briefly noted in the next section.

4. Further Discussion

It is worthwhile to note that

$$L(\mathbf{Y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ \sum_{g=1}^{G} \pi_g N_p(\mathbf{y}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \right\} = \Sigma L(\mathbf{Y} \mid \boldsymbol{\theta}, \mathbf{z}), \tag{14}$$

where the summation of (1), on the right, is over all G^n allocations of the n observations $\mathbf{y}_i$ to the G components. It is the likelihood (14) that is maximized with respect to $\boldsymbol{\theta}$ in Wolfe's reports and by Day (1969). Given the ML estimate of $\boldsymbol{\theta}$, then each of the n observations $\mathbf{y}_i$ is assigned to the component for which $\hat{\pi}_g N_p(\mathbf{y}_i \mid \hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}}_g)$ is largest. With large samples this procedure is optimal and fortunately relieves the concern for searches over G^n allocations to optimize criteria such as those presented in §2. However, these same criteria appear to perform better with small samples and can profitably be used to provide initial estimates, (3), (4), and (5) or (7), for a program to maximize (14), as noted in Symons's 1973 report.

Marriott (1975) has pointed out that such estimates of $\boldsymbol{\theta}$ from a partition $\hat{\mathbf{z}}$ or $\tilde{\mathbf{z}}$ are inconsistent. Clearly the estimates (3), (4), and (5) or (7) are conditional on $\hat{\mathbf{z}}$. Estimates based upon (14) are then unconditional as regards any partition. That these conditional estimates are inconsistent is intuitively reasonable, as one can see that differences between means will be overestimated and variances underestimated whenever there is overlap in the mixture components and estimates are computed, given such partitions of the data. It should be kept in mind, however, that in cluster analysis the interest is in the estimation of the best allocation of the n observations to the G groups, as noted by Scott and Symons (1971), in Symons's 1973 report, and by Binder (1978). If estimation of the parameters of the model (1) is primarily of interest, then the likelihood in (14) is the cornerstone of a theoretically-sound estimation of $\boldsymbol{\theta}$ from a maximum likelihood or a Bayesian perspective.

Acknowledgements

The suggestions of the Associate Editor and the referees were very helpful and are gratefully acknowledged. Appreciation is also extended to Mr Yang Yuan for his expert programing assistance. This research was completed under an Intergovernmental Personnel Act appointment with the Health Effects Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina.

Résumé

On présente de nouveaux critères de classification lorsqu'un mélange de distributions multivariates gaussiennes est adéquat. Ils dérivent des approches du maximum de vraisemblance et bayesien correspondant à différentes suppositions sur les matrices de covariances des composantes du mélange. Deux des critères sont des modifications du critère du déterminant de la somme des carrés entra-groupes de Friedman et Rubin (1967, *Journal of the American Statistical Association* **63,**

1159–1178); ils par
adaptés à des grou
On décrit les as
(1974, *Biometrics* 3
demandant la sépa
ayant une structure
de formes différent
adaptés aux groupe
classification comp

Binder, D. A. (197
Chernoff, H. (197
Symposium o
California Pre
Day, N. E. (1969).
463–474.
Everitt, B. S. (197
Friedman, H. P. ar
American Stat
Geisser, S. (1966)
149–63. New
Geisser, S. and C
Journal of the
Jeffreys, H. (1961)
Lee, K. L. (1979).
708–714.
Lindley, D. V. (19
Inference. Car
Maronna, R. and
Biometrics **30.**
Marriott, F. H. C.
McRae, D. J. (197
Science **16,** 4
Reaven, G. M. an
multidimensic
Rohlf, F. J. (1970
SAS (1979). *SAS*
Scott, A. J. and
Biometrics **27**

1159–1178); ils paraissent être plus sensibles aux différences de taille des groupes. Deux autres sont adaptés à des groupes de taille différente.

On décrit les aspects pratiques de ces critères, et d'un autre étudié par Maronna et Jacovkis (1974, *Biometrics* **30**, 499–505) pour des matrices de covariance hétérogènes. On donne un exemple demandant la séparation de deux types de malades diabétiques de sujets normaux, chaque groupe ayant une structure de covariance distincte. Les résultats avec les trois critères adaptés aux groupes de formes différentes sont comparables et ils sont préférables à deux obtenus pour les trois critères adaptés aux groupes de forme semblable. On présente, pour l'exemple étudié, deux procédures de classification complementaire.

References

Binder, D. A. (1978). Bayesian cluster analysis. *Biometrika* **65**, 31–38.

Chernoff, H. (1970). Metric considerations in cluster analysis. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, 621–629. Berkeley: University of California Press.

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika* **56**, 463–474.

Everitt, B. S. (1979). Unsolved problems in cluster analysis. *Biometrics* **35**, 169–181.

Friedman, H. P. and Rubin, J. (1967). On some invariant criterion for grouping data. *Journal of the American Statistical Association* **63**, 1159–1178.

Geisser, S. (1966). Predictive discrimination. In *Multivariate Analysis*, Krishnaiah, P. R. (ed.), 149–63. New York and London: Academic Press.

Geisser, S. and Cornfield, J. (1963). Posterior distributions for multivariate normal parameters. *Journal of the Royal Statistical Society, Series B* **25**, 368–76.

Jeffreys, H. (1961). *Theory of Probability*. Oxford: Clarendon Press.

Lee, K. L. (1979). Multivariate tests for clusters. *Journal of the American Statistical Association* **74**, 708–714.

Lindley, D. V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint. Part 2: Inference*. Cambridge: Cambridge University Press.

Maronna, R. and Jacovkis, P. M. (1974). Multivariate clustering procedures with variable metrics. *Biometrics* **30**, 499–505.

Marriott, F. H. C. (1975). Separating mixtures of normal distributions. *Biometrics* **31**, 767–769.

McRae, D. J. (1971). MIKCA: a Fortran IV iterative *K*-means cluster analysis program. *Behavioral Science* **16**, 423–424.

Reaven, G. M. and Miller, R. (1979). An attempt to define the nature of chemical diabetes using a multidimensional analysis. *Diabetologia* **16**, 17–24.

Rohlf, F. J. (1970). Adaptive hierarchical clustering schemes. *Systematic Zoology* **19**, 58–62.

SAS (1979). *SAS User's Guide*, 1979 *Edition*. Raleigh: SAS Institute Inc.

Scott, A. J. and Symons, M. J. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics* **27**, 387–397.

Received December 1979; *revised June* 1980

NOTICE

This document has been reviewed in accordance with U.S. Environmental Protection Agency policy and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

This article documents research that was conducted under the auspices of the U.S. Environmental Protection Agency (USEPA). As an arm of the government, the USEPA has a paid up, non-exclusive, irrevocable, worldwide license to reproduce, prepare derivative works, and distribute to the public all of the material contained herein, and to grant others acting on its behalf the right to do so. The USEPA grants to the National Technical Information Service, Department of Commerce, the right to act as its agent in the distribution of this material.